Public access to USA Federally funded scholarly publications

(Response to US OSTP RFI)

Arthur Sale, PhD

Emeritus Professor of Computer Science, University of Tasmania, Australia

Formerly Professor of Computing Research 2002-200), Pro Vice-Chancellor 1993-1999, Chair of Academic Senate 1990, Professor of Computer Science 1974-93.

National Vice-President, Australian Computer Society 1979, 1980 and 1983. Chair Association of Professors of Computer Science 1985-87, etc. ANCAAC Prize 2002, Lifetime Award for contribution to ICt, etc.

[Contact details at end]

Question 1

'Are there steps that agencies could take to grow existing and new markets related to the access and analysis of peer-reviewed publications that result from federally funded scientific research? How can policies for archiving publications and making them publically accessible be used to grow the economy and improve the productivity of the scientific enterprise? What are the relative costs and benefits of such policies? What type of access to these publications is required to maximize U.S. economic growth and improve the productivity of the American scientific enterprise.'

Yes, there are steps that agencies can take. Publishers assert rights over the published forms of articles, which are probably not justifiable (see later), but also often over-ride the dubious claim by requiring transfer of copyright, or enforce restrictive monopoly practices such as embargo periods. As agencies fund the research, any requirements they place on the grantees regarding publications emerging from funded research constitutes a prior legal requirement, binding the grantee and making it illegal for them to sign away these rights.

The basic step any government agency can take is to make it a condition of accepting a grant offer that:

1. Every peer-reviewed publication arising from the grant shall be made open access (in other words freely downloadable to and readable by any person with access to the

Internet and appropriate display software) not later than one month after the date of publication.

2. Acceptable forms of open access include:
   (a) Publishing in a journal which provides open access at publication time.
   (b) Depositing the *Version of Record* of the publication (=VoR; NISO terminology) in a repository and marking it open access (it does not matter if the repository is institutional though this is preferable, or a subject repository).
   (c) Depositing the Version of Record of the publication in a social networking tool such as Mendeley and making it public (open access).
3. When assigning a licence to publish to a publisher, the publisher shall be informed by the grantee of the prior legal constraint agreed to by the grantee. Failure to do so will not invalidate the requirement.
4. It shall be a requirement that in the final acquittal of the grant, the grantee shall list all publications arising from the grant together with their Internet URL (or acceptable alternative such as a DOI) and the date on which it was made open access.

The intention is to provide low-cost open access, with freedom of choice of open access route by the grantee. Spot audits should be carried out on the final acquittal to ensure compliance, and grantees should be informed of this at the time of accepting the grant. Responsibility for these audits could be delegated to the grantees' institutions with appropriate reporting. There should be consequences for non-compliance.

There is one further thing that could be done, though it is not essential. Under present conditions, route 2(a) is limited by the uneven playing field *as seen by the researcher*. As many open access journals charge reader-side fees, authors see this as a charge for a service they have always got for free. They do not see the economics of publishing from a wider perspective. Of course, when the majority of journals have become open access, the institutional funds at present used for subscriptions can be diverted to reader-side fees, but in the meantime, the growth rate of open access journals is limited to be so slow that some authors have predicted 2035 as the year in which this situation may be achieved. The projected date and the assumptions are disputable, but this highlights the problem.

The playing field may be levelled, and the growth in the number of open access journals accelerated, by simply allocating (at the agency level) a small amount of grant funds for paying reader-side fees, capped at say $1,500 per publication to prevent exploitation. These funds might then be allocated to grantees' institutions according to available data, for local administration and disbursement. The advantage of this is that the decision is made at a level where the total cost of the publishing system is evident and used in policy-making. Funds would not be available for so-called 'hybrid' practices which are generally exorbitantly priced, and offer no long term benefit. Any such system of subsidizing researchers for author-side fees should be regarded as strictly transitional and reviewed after say three years.

Routes 2(b) via institutional repositories, and route 2(c) through social networking apps appear free to the researcher anyway. Route 2(b) using a subject repository may be free but the funding responsibility to unclear, but it is undesirable to encourage subject repositories since they create and rigidify silos. Multi-disciplinary research would be hindered.

Question 2

'What specific steps can be taken to protect the intellectual property interests of publishers, scientists, Federal agencies, and other stakeholders involved with the publication and dissemination of peer-reviewed scholarly publications resulting from federally funded scientific research? Conversely, are there policies that should not be adopted with respect to public access to peer-reviewed scholarly publications so as not to undermine any intellectual property rights of publishers, scientists, Federal agencies, and other stakeholders?'

The answer to question 1 addresses all these questions, except the intellectual property rights of publishers. Where exactly this idea arose that publishers created an intellectual property right in an article though copy-editing is not clear. It may be at the time that the Internet and pdf files began to impinge on journal publishers. Intrinsic intellectual property rights from copy-editing are not the norm in the book publishing industry, where I can look at book after book (fiction, non-fiction or textbook) and find on the title page the copyright attributed to the author or authors. If proof-reading, editing, pagination, binding etc do not create intellectual property rights in books, why should they in the case of scholarly journal publishers who exert far smaller effort, make far fewer creative decisions, and take far less risk than a book publisher?

A competent defence lawyer should be able to cite the above and other examples (such as the non-acquisition of IP by typists (in the past) and in-house copy-editors should this be challenged in court. The only cases which are exceptions, and which can be seen in the book publishing industry, are when a publisher or another person is commissioned to add photographs or illustrations to a text-based manuscript (for example see Richard Dawkins 'Climbing Mount Improbable'. Such instances would be rare in the scholarly publishing industry as the authors supply diagrams, photographs and illustrations in the manuscript. They would arise only in the populist science press such as *Scientific American*, *American Scientist*, *National Geographic*, etc, which are assumed to be outside the definition of 'publication' in this RFI.

If intrinsic intellectual property rights of publishers do not exist, their protection is not necessary. If and when authors transfer some copyright or a licence to a publisher, the Copyright Law is perfectly adequate to deal with the acquired rights. The proposed action (see Question 1) simply limits the rights that authors may dispose of, so as to preserve the public benefit of open access.

No steps are necessary.

Question 3

'What are the pros and cons of centralized and decentralized approaches to managing public access to peer-reviewed scholarly publications that result from federally funded research in terms of interoperability, search, development of analytic tools, and other scientific and commercial opportunities? Are there reasons why a Federal agency (or agencies) should maintain custody of all published content, and are there ways that the government can ensure long-term stewardship if content is distributed across multiple private sources?'

This question involves some duplication, but this will be supplied (less succinctly) so the response stands alone. To the two categories mentioned I have added a third: Personal Repositories, which might be considered a limiting case of decentralized repositories, but has especial and different consequences. I do not see this question as central to the RFI. The task force should not be setting up repositories or writing software.

Decentralized Repositories

This term is interpreted to mean institutional repositories, which are not subject-specific. It includes the possibility of a consortium of institutions sharing a repository. Subject-based searches require tools which are readily available.

Pro

- The institution has the legal power to compel its employees (especially researchers) to deposit materials produced by them during the period of employment in its own repository (a 'mandate'). This goes beyond the scope of agency control.
- The institution also has the legal power to compel its graduate students to deposit materials produced by them during their enrolment in its own repository. This is also outside agency control.
- In principle, this implies that institutions can ensure the capture of born-digital versions of all publications produced by their researchers and graduate students. It does ***not*** mean that all such materials will be made open access. This can potentially capture close to 100% of publications in a born-digital form, not just the Federal funded publications.
- The number of repositories is not large in Internet terms, at most the number of research institutions in the USA.
- Internet search requires use of global search tools which aggregate the contents of all websites or repositories known to them. Examples include Google Scholar, Scirus and BASE.
- Preservation of records is the responsibility of the institution, and will be carried out with varying degrees of effectiveness depending on the institution.

Con

- Most researchers are not able to keep copies of all their publications in one repository, as they move institutions.
- There are no tools for discovering new decentralized repositories, not for new repositories to advertise themselves. There are voluntary directories of varying effectiveness.
- The cost of running decentralized repositories is a charge on the institutions. This charge is not large, but it is also not zero.
- The cost of maintaining a mandated deposit regime is a cost to the institutions.
- Researchers identify first with their careers, then with their disciplines, and with their employers last This affects the attitude of researchers to decentralized repositories.

Centralized Repositories

This term is interpreted to mean repositories which are limited to a single subject, and which hold all the works supported by an agency, while perhaps holding others.

Pro

- Many [older] researchers are happy with discipline-based repositories, as they conform more to their pre-Internet ideas. (These ideas are no longer valid and are based on discovery limitations of the print era.)
- There is a small chance of subject-specific tools, perhaps exemplified by searchers for gene names (biology), or crystallographic structures (chemistry). These are relatively little used and can be provided on gateways instead.

Con

- Centralized repositories cast discipline boundaries into concrete and reinforce silos. (Centralized gateways harvesting from decentralized repositories do not, at least to the same extent. They can always be paralleled by another gateway with a different admission strategy.)
- Multi-disciplinary research is inhibited. (Gateways based on discipline repositories will be so broad as to be no better than global search.)
- The responsibility for running the repository is unclear. It may well result in claims for Federal funding.
- Access by international researchers to Federally funded centralized repositories would be unclear.
- There are actually very few centralized repositories, so their adoption as a model involves high set-up costs. Agencies may not be prepared to fund them or guarantee continuance.

Personal Repositories

This term is interpreted to mean software that allows a researcher to create a personal repository of an individual corpus of publications, such as Mendeley. This can be considered to be a limiting case of decentralization, but has special features since it is allied to social networking. I have called it the Titanium Road to OA.

Pro

- Personal repositories utilize the phenomenon of the social media such as Facebook, YouTube, Twitter, etc. Mendeley is cited here as the ‘killer app’, but may be superseded by an even better product under market forces, or by a group of apps (like browsers). They attract high numbers of participants, without compulsion.
- Preliminary data show that the number of researchers in the world is around 3.6M, while Mendeley currently claims 1.4M users. This is about 40% penetration, allowing for some non-researchers to be Mendeley users.
- The annual growth rate in usage is estimated at between 37% and 74% pa.

- Open Access publications on personal repositories is no harder to harvest and aggregate than institutional repositories. The numbers of sites are perhaps 3M instead of 20k, but this is easily do-able in current technology.
- It is relatively easy for publications to be copied from repositories to social apps.
- It is a bit more difficult to migrate publications from social apps to repositories, though one trial is under way (DURA). Full synchronization is probably not wanted by the repositories, since researchers move institutions, and the institutions do not want all their prior publications.
- Social apps may result in multiple copies of a publication becoming OA. This is a net benefit to preservation probability.

Con

- Social apps distract attention from the other two routes to open access (Gold and Green)
- This route poses a bigger risk to the recalcitrant toll-access publishers than any other. Their move to a different business model (for example author-pays) is accelerated. This could be considered as a benefit.
- Discovery of relevant publications is perhaps slightly more difficult, but only slightly.

The adoption of Mendeley has resulted in publications being made OA at a rate at least equal to that of a departmental mandate. It is not known yet, and is a matter of research, to what extent Mendeley users are a disjoint or overlapping set from those providing OA through other routes.

Final Two Questions

There ***no*** compelling reasons why Federal agencies should maintain custody of all published content [funded by them]. There ***are*** compelling reasons for them ***not*** being involved. Obviously taking on such a task involves the expenditure of taxpayer funds, and it creates unnecessary silos, not only between disciplines, but within disciplines. This inhibits multi-disciplinary research, and research into new fields.

The best way to ensure long-term survival of content is to distribute multiple copies of the document across the Internet. Not to excess, but adequately. That is how the scientific literature of the past (and such ancient and medieval manuscripts as survived) did. There is scope for some Federal funds to be applied to making interoperability between open access journals, decentralized repositories, centralized repositories, and personal repositories as easy as possible, as well as strongly recommending that all of these adhere to agreed standards.

This implies that if an open-access journal becomes bankrupt and perchance its online presence is destroyed, its footprint is still largely or completely visible on the Internet. Current search engines can easily deal with duplicates, and it will only become easier and easier. If any repository closes (or its host institution), the content will survive. Personal repositories may not long survive the death of their owner, but the Library of Congress may wish to advise on whether any action should be taken in respect of them.

There is one facet of this advice which should be drawn out. It would be desirable for *all* copies of a document online to have the same ancestor. There can be no doubt that this needs to be the *Version of Record*, or in other words the version that was published, as an electronic file. Its form may be xml, html, pdf or another form – it does not matter much to ICT specialists. It is undesirable for multiple versions to be simultaneously online (such as a *Draft Manuscrip*t or the *Accepted Manuscript*) unless they are attached to the primary document as prior versions and clearly identified as such. In some cases, authors may want to attach a *Long Version* or an *Extended Version* (not NISO terms).

Question 4

'Are there models or new ideas for public-private partnerships that take advantage of existing publisher archives and encourage innovation in accessibility and interoperability, while ensuring long-term stewardship of the results of federally funded research?'

I submit three new cooperations:

1. Funding or supporting gateway systems which provide access to aggregations of the data stored in the cloud, but perhaps provide discipline, subject or emerging interest foci. This does not involve migration of data, but search and discovery.
2. Funding or encouraging standardization and interoperability between ***all*** the forms of open access to scholarly literature (persona; plain html web pages perhaps excluded). See above re creating multiple copies on the internet.
3. Supporting a Library of Congress program to monitor and address what to do when the number of open access copies of a federally funded scholarly publication on the Internet drops close to extinction (say one copy). There might be a Repository of Last Resort, for example. At the start, toll-access journals with the usual monopolistic practices will need to be treated as special cases, because they may only permit single (or no) copies to be open access online (theirs), against the public interest.

Question 5

'What steps can be taken by Federal agencies, publishers, and/or scholarly and professional societies to encourage interoperable search, discovery, and analysis capacity across disciplines and archives? What are the minimum core metadata for scholarly publications that must be made available to the public to allow such capabilities? How should Federal agencies make certain that such minimum core metadata associated with peer-reviewed publications resulting from federally funded scientific research are publicly available to ensure that these publications can be easily found and linked to Federal science funding?'

This question is trivial. The standards for metadata already exist, though there are arguments at the margin. All that is needed is to pull together a common core standard. If interoperability of all sources is a key part of the outcome of this RFI, compliance will follow swiftly.

As to discovery, this is largely both trivial and an insoluble problem. Search facilities like Google Scholar and Scirus make a better fist of discovery than any publisher, agency or society can hope to achieve. Let them get on with the task. Precision *vs* coverage is an

insoluble problem, though helped by current research by computer scientists into analysing personal needs and predicting useful outcomes. Again, nothing is needed from the Federal government, apart from its usual research grants.

If one knows the citation of an article it is of course totally trivial to find it on the Internet if it exists. Simply search for the title as a literal string, and search for an author with an unusual name, or several authors. The result is almost guaranteed to be definitive.

Metadata are not very important in this discovery process. They mainly help gateways and other secondary services, and automatic metadata extraction is almost as good (if not better than) librarian custom metadata.

Question 6

'How can Federal agencies that fund science maximize the benefit of public access policies to U.S. taxpayers, and their investment in the peer-reviewed literature, while minimizing burden and costs for stakeholders, including awardee institutions, scientists, publishers, Federal agencies, and libraries?'

This is a strange question. It is really two questions because the link between the halves is tenuous. The maximization is easy: maximize free, open access worldwide, thereby both encouraging maximization of access to US residents, and international collaborations. Also use this policy as a lever to get other countries to reciprocate. Australia should be easy, if a little pressure is applied, as a long-standing ally of the USA.

The minimization is addressed in the rest of this response to the RFI. Awardee institutions have minimal obligations but receive research funding anyway; scientists have minimal burdens; publishers are impacted but mainly to make them give up monopoly practices and change their business model to a sustainable one (sooner rather than later for US-reliant publishers one hopes), Federal agencies have minimal obligations, and libraries simply absorb the changing world into their practices.

There is no need for any special action.

Question 7

' Besides scholarly journal articles, should other types of peer-reviewed publications resulting from federally funded research, such as book chapters and conference proceedings, be covered by these public access policies?'

Articles in conferences should certainly be included, because they are key forms of publications in ICT (information and communication technologies) which is a key driver in industrial growth (including genetics and biological science). Conference *proceedings*, as a collection, need not be included as a result of this RFI round. They are too easy to confuse with books.

Book chapters and conference proceeding should perhaps be left for a later time, after the journal and conference article open access is firmly embedded in researcher and publisher

consciousness. Books themselves often involve royalties (which scientific articles don't) in exchange for publication rights.

Question 8

'What is the appropriate embargo period after publication before the public is granted free access to the full content of peer-reviewed scholarly publications resulting from federally funded research? Please describe the empirical basis for the recommended embargo period. Analyses that weigh public and private benefits and account for external market factors, such as competition, price changes, library budgets, and other factors, will be particularly useful. Are there evidence-based arguments that can be made that the delay period should be different for specific disciplines or types of publications?'

Simple question, though an irrelevant one. Publishers have no rights to set an embargo period unless researchers assign such a right to the publisher. The optimum period is zero months.

Each extra week embargo means lost impact of the research and damage to taxpayer return. This is evidenced by the citation data, which are themselves simply lagging indicators of impact.

There are no arguments for different discipline embargo periods. Zero means exactly zero days after publication and zero times any factor is still zero.

It should be emphasized that publishers have no rights to set embargo periods, except by exercising bullying power and pseudo-legal claims, unless researchers (the basic copyright owners) assign them these rights. They should not be allowed to, through the prior legal status of the grant acceptance contract.

Any Other Matter

'Please identify any other items the Task Force might consider for Federal policies related to public access to peer-reviewed scholarly publications resulting from federally supported research.'

Publisher rights

I have referred to the opinion that publishers have no *intrinsic* legal claim to any rights over journal articles arising from copy-editing or preparation for publishing. It is worth teasing this out a bit more.

The institutional mandate idea has been publicized as ID/OA: In other words 'Immediate Deposit, Open Access when possible'. This attitude is based on not confronting publishers with their claim to implicit copyrights, and requires a researcher to immediately deposit their *Accepted Manuscript* (the final pre-publication draft), and to make it open access. Unfortunately, while readers of articles are happy to read an AM instead of a VoR, researchers are not happy to deposit an AM and frequently do not, preferring to deposit a VoR and make it Restricted Access.

As suggested in this submission, the journal publishers' claim to intrinsic copyrights derived from their copyediting and publishing activities are spurious. This is why the publisher lobby have sought to enact a US law: *Research Works Act (H.R.3699)* to create such a new right. This is totally unjustifiable. The proposed Act should be rejected; it is unlikely to be accepted by other countries as it is contrary to the common copyright laws of the world, based on the Berne Convention.

Researcher attitudes

It is clear that prior to publication, most researchers are chary of confronting publishers with any quibbles, and they sign whatever they are presented with as a contractual agreement. Researchers are eager for their work to be published, whatever that means. They have little knowledge of law, in general, nor do they have agents who are.

After publication, the situation changes. Researchers have always regarded the article they authored as 'theirs', and they treat the Version-of-Record as theirs also. Legally, they are correct, unless they signed away all their rights. If anyone asks for a copy of their publication, they dispense the VoR as a matter of routine, whether in times past this was a reprint or a photocopy, or now it is a pdf file. This applies across the spectrum of disciplines, though there are some slight discipline variations.

In addition, researchers treat publications of which they are not an author as in the public domain, once they have them in their hand. It is the VoR that they identify as the publication. Seldom does a week go by on any of the mailing lists to which I subscribe that I do not receive something like the following (de-identified, but not otherwise edited):

> Hi there.
>
> Someone can help me to get a copy of the following paper?:
>
> A rapid and inexpensive microplate assay for the enzymatic determination of glucose, fructose, sucrose, L-malate and citrate in tomato (Lycopersicon esculentum) extracts and in orange juice Joyce S. Velterop, Femke Vos Phytochemical Analysis Volume 12, Issue 5, pages 299–304, September/October 2001
> DOI: 10.1002/pca.598
>
> Thanks in advance

Followed the next day by the post:

> Thank you so much!!!

This week I received notice of at least three such exchanges. It is clear that as far as researchers are concerned, the copyright law might as well not be enacted. Scholarly articles that they or others wrote are considered to be in the public domain. Publishers are acquiescent in this attitude, and do not contest any transfers of this kind, though they must be aware of them.

Minimal intervention

This submission suggests a minimal intervention route for the Task Force.

The proposed policy simply interposes a condition of accepting a grant (that it be made OA and will be audited) and leaves the route of achieving this to the grantee and their institution.

- Market forces can determine which form of OA is effective or dominant. It does not matter, as long as the article is OA. This is consistent with general US attitudes.
- Every researcher has at least one free solution, being the social network access exemplified by Mendeley.
- The articles on the Internet are all of the canonic form of the published version, eliminating confusion, and fitting with researcher wishes.
- Publisher claims to copyright which are unsustainable are rejected. However, publishers are also provided with market force freedom. It is open to them to transition to open access journals (often author-side fees), or to provide selective (hybrid) OA to publications funded by Federal Government agencies.
- The audit requirements are simple, low cost, and largely devolved to grant-recipient institutions. However, agencies need to enforce penalties for non-compliance (like ineligibility for an institution to receive grants for say three years), which should be easy.
- Publishers and publisher associations should be informed that researchers in receipt of Federal grants are unable to dispose of copyright without securing an immediate open access right for their articles. Agreements signed by the author contrary to this will be invalid. Agencies should be prepared to challenge and contest in a court of law any author agreement which does not comply with the proposed requirement.

Outside these, the open access transition details are left to private enterprise and market forces. Publishers will adapt or adopt new business models. Some may offer free open access routinely to grant recipients (a hybrid model), while others may transition to full open access journals. A minimalist solution of this kind should be highly attractive to US citizens.

References

Few references are quoted, because otherwise they would be voluminous and detract from the submission. However, I offer the following few important references.

NISO (2008). 'NISO RP-8-2008, Journal Article Versions (JAV): Recommendations of the NISO/ALPSP JAV Technical Working Group'. http://www.niso.org/publications/rp/RP-8-2008.pdf .

Sale, AHJ (2011) The Titanium Road Essays. Discussion Paper, Hobart, Tasmania, Australia. http://eprints.utas.edu.au/12526/.

Sale, AHJ (2006) The acquisition of open access research articles. First Monday, 11 (10). http://www.firstmonday.org/issues/issue11_10/sale/index.html.

Harnad, S and Carr, L and Swan, A and Sale, AHJ and Bosc, H (2009) Open Access Repositories - maximizing and measuring research impact through university and research-funder open-access self-archiving mandates. Wissenschaftsmanagement, 4 (4). pp. 36-41. ISSN 0947-9546. http://eprints.utas.edu.au/9273/.

Contact Details

Emeritus Professor Arthur Sale, PhD
Locked Bag 100
School of ICT
University of Tasmania
Tasmania, Australia 7001

Email: Arthur.Sale@utas.edu.au (university), or ahjs@ozemail.com.au (home)

Fixed phone: +61-3-6247 1331
Cell phone: +61-4-1947 1331

Skype:Arthur.Sale